Adaptive Optimal Control for Time-Varying Discrete-Time Linear Systems

Shuzhi Sam Ge[1], Chen Wang[1] and Yanan Li[2]

Abstract: In this paper, adaptive optimal control is proposed for time-varying discrete linear system subject to unknown system dynamics. The idea of the method is a direct application of the Q-learning adaptive dynamic programming for time-varying system. In order to derive the optimal control policy, a actor-critic structure is constructed and time-varying least square method is adopted for parameter adaptation. It has shown that the derived control policy can robustly stabilize the time varying system and guarantee an optimal control performance at the same time. As no particular system information is required throughout the process, the proposed techniques provide a potential feasible solution to a large variety of control application. The validity of the proposed method is verified through simulation studies.

Index Terms – adaptive dynamic programming; adaptive optimal control; time-varying linear discrete system

I. Introduction

All control problems involve regulating a system's input so that the system can meet some desired specifications. Among all the design criteria, stability is the most fundamental requirement. Controllers that can robustly stabilize a particular set of plant have already been studied in many research works [1], [2], [3], [4], [5]. The development of modern industry has raised a higher goal which require the system can be stabilized with small amount of energy. This higher goal is generally considered as the optimality of a system which is explained as "the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."[6].

In the literature of optimal control, two approaches are most widely studied, , i.e., classical optimal control based on maximum principle [7], [8] and dynamic programming [6]. In classical optimal control, a finite or infinite cost function is defined to describe the control performance, and control policy is generated by solving the famous Algebra Riccati Equation (ARE) where the state-space model of the system is assumed to be known. In dynamic programming [6], the optimal solution is acquired by solving a sequence of partitioned problems, the characteristic of dynamic programming lies the multistage nature of the optimization procedure.

However, in those conventional optimal control methods, the system information is assumed to be known, which indicates that the control engineer needs to make an effort to identify the system model in order to build an optimal controller. This process is usually quite tedious and sensitive to system uncertainties and disturbance, making it not desirable in real application. In addition, considering the time varying nature of most system plant model, the traditional methods are not practical in most situations due to the off-line optimization and slow response to plant parameter variations. To tackle this problem, adaptive dynamic programming (ADP) or actor-critic learning is proposed in [9], [10]. ADP mimics the way that biological systems interact with the environment. In the scheme of ADP, the system is considered as an agent which modifies its action according to the environment stimuli. The action is strengthened (positive reinforcement) or weakened (negative reinforcement) according to the evaluation of a critic. By using ADP, an optimal control policy can be generated with partial or none information of the system. This is a heuristic process where an agent tries to maximize its future rewards; in a control engineering context, the maximization of reward is equivalent to the minimization of a control cost. Among all these ADP approaches, most recognized discrete ADP algorithms are the heuristic dynamic programming (HDP), globalized DHP (GDHP), action-dependent heuristic dynamic programming (ADHDP) or Q-learning [11] and dual-heuristic programming (DHP). The common feature of these ADP algorithms is that the design of optimal controller only requires partial information of the system model to be controlled.

Among all these schemes, for discrete-time systems, ADHDP[12], or Q-learning [13], is an online iterative learning algorithm which does not rely on the specific plant model to be controlled. Due to its unique online learing and control structure. This method has been widely applied in many research fields, such as nonzero-sum games [14], robotic arm control [15] and optimal output feedback control designs [16].

The idea of the proposed adaptive optimal control method is similar to [14]. However, instead of developing a adaptive optimal control for a time invariant system, we further considered the time-varying nature of actual plant model and try to solve the following difficulties, i.e., (a) the time-varying system parameters are not computationally tractable in real practice, and (b) when the parameters of the plant

[1]Shuzhi Sam Ge and Chen Wang are with the Social Robotics Lab, Interactive & Digital Media Institute (IDMI) and the Department of Electrical & Computer Engineering, National University of Singapore, Singapore 117576 samge@nus.edu.sg

[2]Yanan Li is with NUS Graduate School for Integrative Sciences and Engineering (NGS), National University of Singapore, Singapore 117456 liyanan84@nus.edu.sg

model change over time, steady-state optimized solutions may be inappropriate. (c) in some special cases (model fault or system failure), the plant model may undergo a suddenly changed during the operation. In these scenarios, common solution methods such as those used in [14] can be used by resetting the controller and updating the control policy based on a new batch of data which may be too conservative. The execution of the optimal policy will then be delayed and slow to such changes.

To address these problems, a more feasible method is proposed in this paper to extend the decision and control structure by including the time-varying parameter, such that for each adaptation step, a different set of parameters and control policy are implemented. Under this scheme, the decision making and policy updating can be done with little computation cost, making it more advantageous.

Based on the above discussion, we highlight the contributions of this paper as follows

(i) The time-varying system model is considered completely unknown for the adaptive optimal control design. The optimal control policy is generated based on the policy iteration.
(ii) Recursive time-varying least square is adopted to derive the optimal control policy which is feasible for smooth online adaptation.
(iii) Two general case of time-varying model are considered in the simulation, which verify the effectiveness of the proposed methods. The simulation results have further proved the system's robustness subject to parameter variation and model uncertainties.

The rest of the paper is organized as follows. In Section 2, the problem is formulated and Q-function for time-varying system is described. In Section 3, adaptive optimal control is developed for the described time-varying plant model and the optimal policy is achieved subject to unknown system model. In Section 4, the validity of the proposed method is verified through simulation studies. Section 5 concludes this paper.

II. Problem Formulation and Preliminaries

A. System Description

Let us consider a linear time-varying discrete system as

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) \end{aligned} \tag{1}$$

where k denotes the time instant, $x(k) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m$, and $A(k)$, $B(k)$ and $C(k)$ are time-varying matrices which are stabilizable.

The optimal control problem can be formulated by designing a controller in the following form

$$u(k) = -L(k)x(k) \tag{2}$$

which minimizes the infinite control performance

$$J = \sum_{k=k_0}^{\infty} [x^T(k)S(k)x(k) + u^T(k)R(k)u(k)] \tag{3}$$

where $S(k) \in \mathbb{R}^{p\times p}$ and $R(k) \in \mathbb{R}^{r\times r}$ are the weights of the state and the input which satisfy $S(k) = S(k)^T \geq 0$ and $R(k) = R(k)^T > 0$, and $L(k)$ is the control gain.

In classical optimal control, if the system information ($A(k)$ and $B(k)$) is completely known. The optimal control policy $u(k)$ can be obtained by solving the following discrete algebraic riccati equation (DARE)

$$\begin{aligned} P(k) = A^T(k)P(k+1)A(k) + S(k) - A^T(k)P(k+1)B(k) \\ [R(k) + B^T P(k+1)B(k)]^{-1}B^T(k)P(k+1)A(k) \end{aligned} \tag{4}$$

The optimal feedback gain $L(k)$ can be further derived by

$$L(k) = [R(k) + B^T(k)P(k+1)B(k)]^{-1}B^T(k)P(k+1)A(k) \tag{5}$$

Remark 1: Due to the time-varying nature and nonlinearities of most plant models, it is not feasible to use this off-line methods in real application scenario. In common practice, close approximation for the optimal solution to the DARE is performed by using system parameters valid at the last time step and replaced by a steady-state solution. This method can ease the numerical computation cost to some extent, however, because this method still relies on the assumption that the time varying matrix $A(k)$ and $B(k)$ are known, which prevents it from further application.

B. Optimal Principle and Q-function for Time-varying LQR problem

As discussed in the previous section, the conventional off-line design of optimal controller is usually time-consuming and suffers from the slow response to parameter variations. To solve this problem, in the following, we will show how to describe and model the problem using the Bellman's principle of optimality and derive an online policy using the concept of Q-functions [12], [13].

Let us consider the following infinite horizon value function of the plant model in (1).

$$V^*(x(k)) = \min_{u(k)} \sum_{k=k_0}^{\infty} [x^T(k)S(k)x(k) + u^T(k)R(k)u(k)] \tag{6}$$

Then our goal is to decide the optimal control policy $u^*(k)$. Assuming the problem has a solvable solution, then it is well known that the cost value $V^*(x(k))$ is quadratic in the state with the following form

$$V^*(x(k)) = x(k)^T P(k)x(k) \tag{7}$$

where $P(k)$ is a time varying matrix. The cost-to-go cost

function can be defined as

$$\begin{aligned} & V^*(x(k)) \\ = \; & g(x(k),u(k)) + V^*(x(k+1)) \\ = \; & x(k)^T S(k) x(k) + u(k)^T R(k) u(k) + \\ & x^T(k+1) P(k+1) x(k+1) \\ = \; & \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}^T \begin{bmatrix} S(k) & 0 \\ 0 & R(k) \end{bmatrix} \begin{bmatrix} x(k) \\ u(k) \end{bmatrix} + \\ & \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}^T \begin{bmatrix} A^T(k) \\ B^T(k) \end{bmatrix} P(k+1) \begin{bmatrix} A^T(k) \\ B^T(k) \end{bmatrix}^T \begin{bmatrix} x(k) \\ u(k) \end{bmatrix} \\ = \; & \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}^T H(k) \begin{bmatrix} x(k) \\ u(k) \end{bmatrix} \end{aligned} \quad (8)$$

where $g(x(k),u(k)) = x^T(k)S(k)x(k) + u(k)^T R(k)u(k)$ is the utility function during the k-th step. $H(k)$ in Equ. (8) can be further written as

$$H(k) = \begin{bmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{bmatrix} \quad (9)$$

Where $H_{xx} = A^T(k)P(k+1)A(k) + S(k)$, $H_{xu} = H_{ux}^T = A^T(k)P(k+1)B(k)$ and $H_{uu} = B^T(k)P(k+1)B(k) + R(k)$ The optimal control policy can be acquired by

$$\begin{aligned} u(k) &= -L(k)x(k) \\ &= -\frac{\partial V^*(x(k))}{\partial u(k)} \\ &= -H_{uu}^{-1} H_{ux} x(k) \end{aligned} \quad (10)$$

Eqs.(9) and (10) are the main equations needed to obtain the optimal control policy. In this paper, the H matrix is referred to as H parameters. Note that if H can be obtained using an online identification method, the plant model dynamics will no longer be needed. In the following, we will show how to describe and model the optimal control problem using the Q-function based optimal principle, which will be further used to approximate the ARE equation solution online later.

Let us define the following state and action based Q function.

$$Q^*(x(k),u^*(k)) = V^*(x(k)) \quad (11)$$

Set $P_0 = 0$, for i=1,2,3,... The optimal control problem described in (9) then becomes finding the optimal control policy $u^*(k)$, which satisfies the following time-varying temporal difference equation

$$\begin{aligned} Q^*(x(k),u^*(k) = g(x(k),u^*(k)) + \\ Q^*(x(k+1),u^*(k+1)) \end{aligned} \quad (12)$$

III. Adaptive Optimal Control for Time-varying Discrete System

In the following, we will show how to solve the temporal difference function (12) using a recursive time-varying least square methods. The existing Q-function $Q^*(x(k),u(k))$ from k-th iteration to ∞ can be parameterized in the following form

$$\begin{aligned} & Q^*(x(k),u(k)) \\ = \; & z(k)^T H(k) z(k) \\ = \; & (z(k)^T \otimes z(k)) \text{vec}(H(k)) \\ = \; & (\text{vec}(H(k)))^T (z(k) \otimes z(k)) \end{aligned} \quad (13)$$

where $z(k) = [x^T(k) \; u^T(k)]^T$. Similarly, the cost function from $(k+1)$-th iteration to ∞ can be derived as

$$\begin{aligned} & Q^*(x(k+1),u(k+1)) \\ = \; & z(k+1)^T H(k+1) z(k+1) \\ = \; & (z(k+1)^T \otimes z(k+1)^T) \text{vec}(H(k+1)) \\ = \; & (\text{vec}(H(k+1))^T (z(k+1) \otimes z(t+T)) \end{aligned} \quad (14)$$

If we define $\hat{z}(k) = (z(k)^T \otimes z(k)^T)$ and $\hat{h}(k) = vec(H(k))$, the temporal difference equation in Eq. (12) then becomes

$$\begin{aligned} Q^*(x(k),u(k)) &= g(x(k),z(k))) + Q^*(x(k+1),u(k+1)) \\ \hat{h}^T(k)\hat{z}(k) &= g(x(k),z(k)) + \hat{h}^T(k+1)\hat{z}(k+1) \end{aligned} \quad (15)$$

During the sampling time interval T, it can be assumed that $\hat{h}(k) \approx \hat{h}(k+1)$, then we have the following linear-in-parameter (LIP) form

$$\begin{aligned} g(x(k),z(k)) &= \hat{h}^T(k)(\hat{z}(k) - \hat{z}(k+1)) \\ &= \theta^T(k)\phi(k) \end{aligned} \quad (16)$$

where $\theta(k) = \hat{h}^T(k)$ is the vector of system dynamic parameter and $\phi(k) = (\hat{z}(k) - \hat{z}(k+1))$ is the regressor vector. The above equation is important as it allows us to optimize over the current control policy by working backward in time. The Q learning algorithm can be regarded as the desired target function we need to approximate $V^*(x(k))$ in the least square sense.

In order to identify the time-varying parameter $\theta(k) = \hat{h}^T(k)$, Recursive Exponentially Weighted Recursive Least Squares(REWRLS) discussed in [17] is implemented in this paper. The REWRLS method are employed to optimize the following block-wise Mean Squared Error (MSE) cost function

$$V[\theta(k),k] = \frac{1}{2}\sum_{i=1}^{k} \lambda^{k-i}(g(x(k),z(k)) - \theta^T(k)\phi(k)) \quad (17)$$

where λ is a parameter such that $0 < \lambda < 1$. The parameter λ is called the forgetting factor. The most recent data is given unit weight, but data that is n time units old is weighted by λ^n. Particularly, small values for λ puts greater emphasis on the recent data. The parameter $\theta(k)$ which minimizes Eq. (17) is given recursively by

$$\begin{aligned} \hat{\theta}(k+1) = \hat{\theta}(k) + K(k+1)(y(k+1) \\ -\phi^T(k+1)\hat{\theta}(k)) \end{aligned} \quad (18)$$

where $W(k)$ is the covariance matrix at time instant k and

$K(k)$ is the estimator gain matrix with

$$\begin{aligned} K(k+1) &= W(k+1)\phi(k+1) \\ &= W(k)\phi(k+1)(\lambda I + \phi^T W(k)\phi(k+1))^{-1} \\ W(k+1) &= (I - K(k+1)\phi^T(k))W(k)\lambda \end{aligned} \quad (19)$$

To avoid $W(k)$ becoming too close to singularity, the covariance matrix is reset as follows

$$W(k) = \rho_0 I, \text{if } \lambda_{\min} \leq \rho_1 \quad (20)$$

The following persistent excitation condition needs to be met to ensure the parameter convergence,

$$\delta_1 I \leq \frac{1}{\lambda}\sum_{i=1}^{\lambda} \phi_{k-1}\phi_{k-1}^T \leq \delta_0 I \quad (21)$$

where $\delta_1 \leq \delta_0$, δ_0 and δ_1 are positive scalars. To address this problem, the exploration noise is added in the input during the parameter adaptation

$$u_e(k) = -K(k)x(k) + e(k) \quad (22)$$

where $e(0, \sigma^2)$ is the zero-mean white noise.

IV. Simulations

In this section, two kinds of time-varying systems are considered to testify the effectiveness of the proposed method, which represent a large variety of time-varying discrete linear system. In the simulation, the weight matrices in (6) are given by $S = [2\ 0; 0\ 2], R = 1$ and operation sampling period is selected as $T = 0.001s$

As the plant model is known in the simulation, the exact optimal feedback gains can be obtained by solving the DARE in (4) which is referred to as "LQR", and compared with the the proposed method in this paper which is referred to as "Proposed". It is necessary to emphasize that the plant dynamics are only available in the simulation and they are usually unknown or need to be estimated in real applications. This is the motivation of this paper, which has already been discussed in the Introduction.

A. System with Sudden Model Shift

In the first case, the plant is initially given by

$$x(k+1) = \begin{bmatrix} 0 & 1 \\ -0.15 & -0.2 \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 0.2 \end{bmatrix} u(k) \quad (23)$$

but for $t \geq 2$, the plant parameter suddenly change so that the plant model is given by

$$x(k+1) = \begin{bmatrix} 0 & 2 \\ -0.2 & -1 \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(k) \quad (24)$$

The simulation results are shown in Figs. 1, 2, 3, 4 and 5. In Fig. 1, the control gains using the proposed methods and the desired control gains using LQR are shown and compared. It is found that the obtained optimal control gain using the proposed methods can accurately track the desired values. More details can be found in Fig. 3, where the H parameters convergence is shown. Fig. 5 demonstrates the state trajectories of the adaptive optimal controller at the initial stage when the initial values are selected as $x(1) = 2$ and $x(2) = -1$. The trajectory of the control inputs is shown in Fig. 2. The simulation results show the system can be robustly stabilized even subject sudden parameter change using the proposed method.

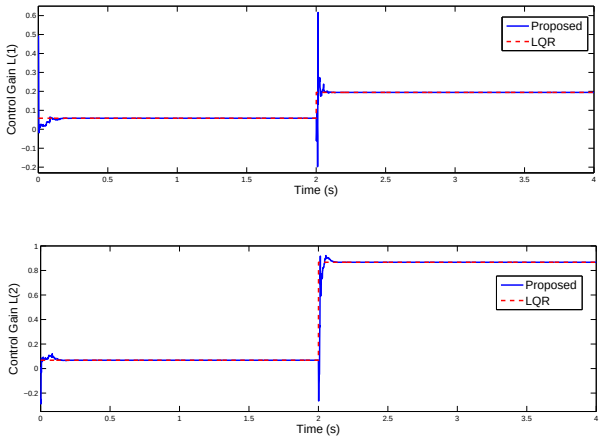

Fig. 1. Desired control gains and actual control gains

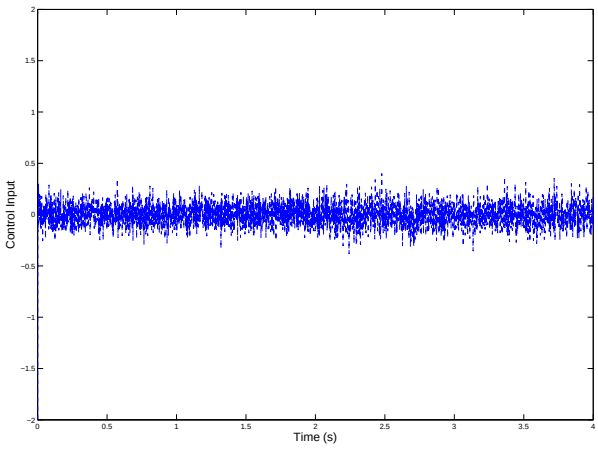

Fig. 2. Control input

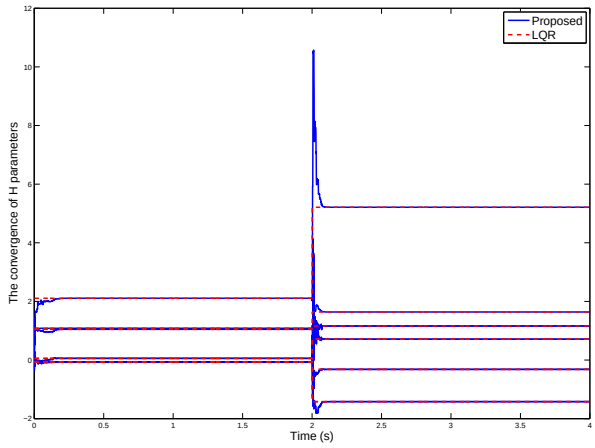

Fig. 3. Convergence of H parameters

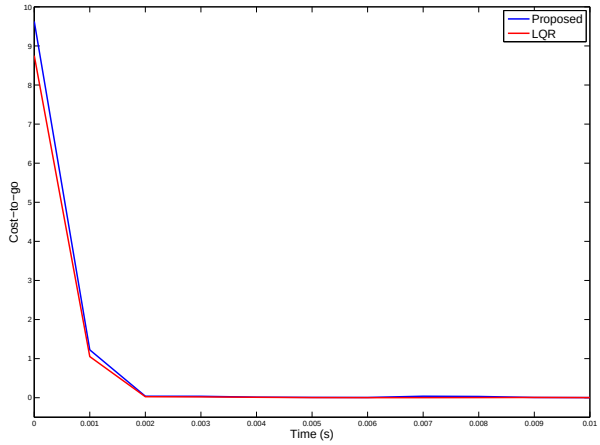

Fig. 4. Cost-to-go

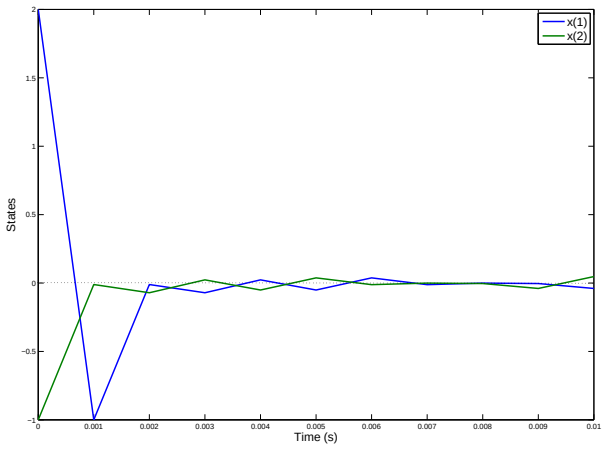

Fig. 5. States trajectories

B. General Linear Time-varying system

In this subsection, the plant is assumed to be a general linear time-varying discrete system which is described by the following state space model

$$x(k+1) = \begin{bmatrix} 0 & 1 \\ -0.2\sin(1.2t) & -0.4\cos(3t) \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 0.2e^{-0.01t} \end{bmatrix} u(k) \quad (25)$$

The initial conditions are the same as in the previous subsection and simulation results are shown in Figs. 6, 7, 8, 9 and 10. Descriptions of the simulation results are similar to the previous section and thus are omitted. From the simulation results, we can conclude that smooth optimal control performance for the general linear time-varying system can be guaranteed using the proposed method.

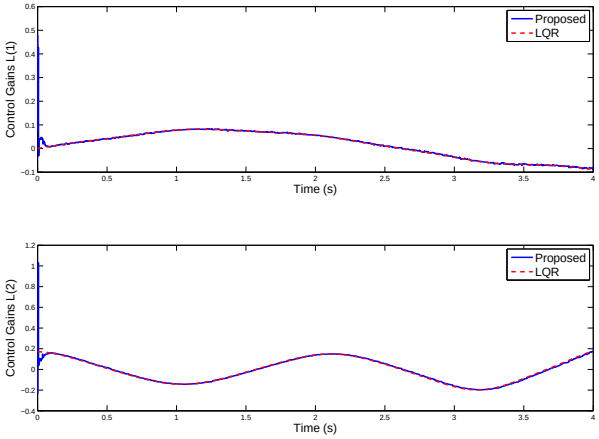

Fig. 6. Desired control gains and actual control gains

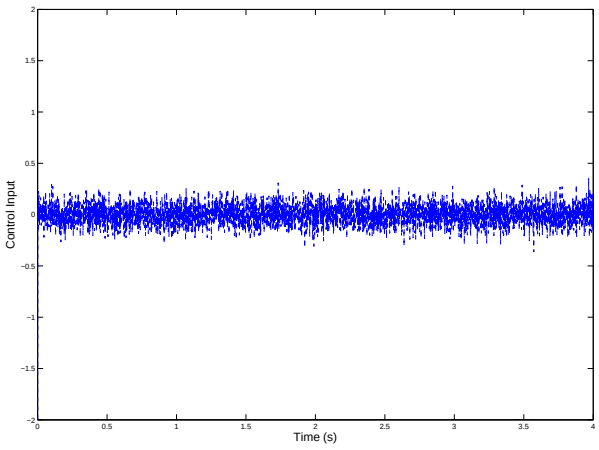

Fig. 7. Control input

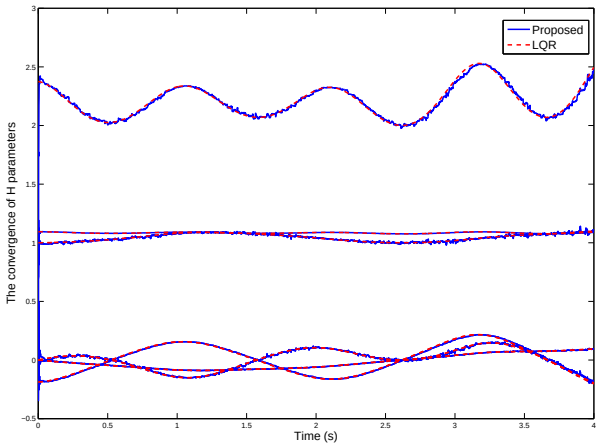

Fig. 8. Convergence of H parameters

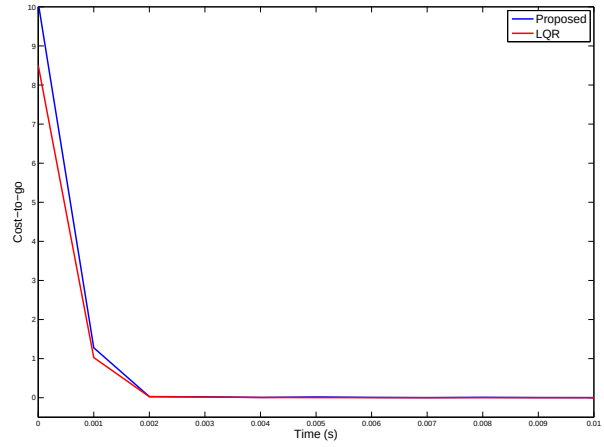

Fig. 9. Cost-to-go

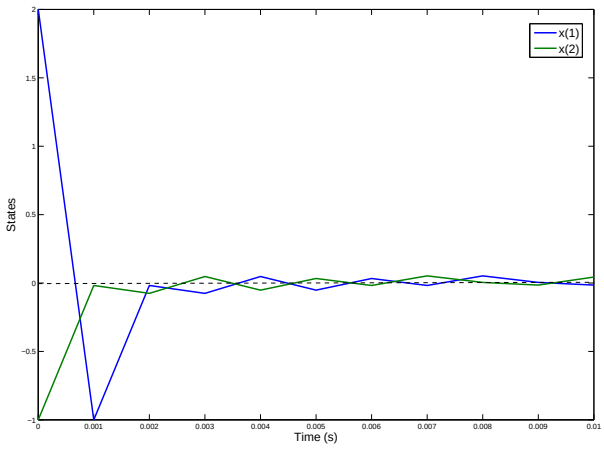

Fig. 10. States trajectories

V. Conclusion

In this paper, an adaptive optimal control is demonstrated for unknown time-varying discrete system. The proposed method is based on Q-learning ADP and similar to [13], [14], but instead of traditional Q-learning for time-invariant system, we further considered the time-varying nature of the system plant and a modified temporal difference equation is employed and solved using REWRLS without requiring any system dynamic information. The simulation for two types of representative time-varying system has further verified the feasibilities of the proposed method.

References

[1] K. Zhou and J. C. Doyle, *Essentials of robust control*, vol. 104. Prentice Hall Upper Saddle River, NJ, 1998.

[2] R. C. Dorf, *Modern control systems*. Addison-Wesley Longman Publishing Co., Inc., 1991.

[3] F. L. Lewis and E. Kamen, "Applied optimal control and estimation," *IEEE Transactions on Automatic Control*, vol. 39, no. 8, pp. 1773–1773, 1994.

[4] G. C. Goodwin, S. F. Graebe, and M. E. Salgado, *Control system design*, vol. 240. Prentice Hall Upper Saddle River, 2001.

[5] S. S. Ge, C. C. Hang, T. H. Lee, and T. Zhang, *Stable Adaptive Neural Network Control*. Norwell, USA: Kluwer Academic, 2001.

[6] R. Bellman and R. E. Kalaba, *Dynamic programming and modern control theory*. Academic Press New York, 1965.

[7] V. BOLTYANSKIY, R. Gamkrelidze, Y. MISHCHENKO, and L. Pontryagin, "Mathematical theory of optimal processes," 1962.

[8] J. Willems, "Least squares stationary optimal control and the algebraic riccati equation," *IEEE Transactions on Automatic Control*, vol. 16, no. 6, pp. 621–634, 1971.

[9] P. J. Werbos, "Using ADP to understand and replicate brain intelligence: The next level design," *Proceedings of the 2007 IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pp. 209–216, 2007.

[10] P. J. Werbos, "Intelligence in the brain: A theory of how it works and how to build it," *Neural Networks*, vol. 22, no. 3, pp. 200–212, 2009.

[11] C. Watkins, "Learning from delayed rewards," *Cambridge University, Cambridge, England, Doctoral thesis*, 1989.

[12] P. J. Werbos, "Consistency of hdp applied to a simple reinforcement learning problem," *Neural Networks*, vol. 3, no. 2, pp. 179–189, 1990.

[13] A. G. Barto, R. S. Sutton, and C. J. Watkins, "Learning and sequential decision making," in *Learning and computational neuroscience*, Citeseer, 1989.

[14] A. Al-Tamimi, F. Lewis, and M. Abu-Khalaf, "Model-free q-learning designs for linear discrete-time zero-sum games with application to h-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[15] S. G. Khan, G. Herrmann, F. L. Lewis, T. Pipe, and C. Melhuish, "A novel q-learning based adaptive optimal controller implementation for a humanoid robotic arm," in *World Congress*, vol. 18, pp. 13528–13533, 2011.

[16] H. Zhang, F. L. Lewis, and A. Das, "Optimal design for synchronization of cooperative systems: state feedback, observer and output feedback," *IEEE Transactions on Automatic Control*, vol. 56, no. 8, pp. 1948–1952, 2011.

[17] K. J. Astrom and B. Wittenmark, *Adaptive Control*. Reading, Mass: Addison-Wesley, 1989.